White Paper Report

Report ID: 106262

Application Number: HD5161812

Project Director: Matthew Knutzen (mattknutzen@nypl.org)

Institution: New York Public Library

Reporting Period: 5/1/2012-7/31/2013

Report Due: 1/17/2014

Date Submitted: 1/21/2014

The New York Public Library

National Endowment for the Humanities
NYC Chronology of Place
White Paper
Grant HD- 51618-12

Background

In April 2012, the National Endowment for the Humanities (NEH) awarded a $50,000 grant to The New York Public Library (NYPL) to build *NYC Chronology of Place* (http://dev.nypl.gazetteer.in/), an experimental Linked Open Data gazetteer. This unique online database enables researchers to connect historical geographic places (e.g. landmarks, buildings, and districts) to fixed locations, and use the results to enhance their work.

Overview

Gazetteers are dictionaries of place names which, when on the web, act as location databases; services like Google Maps rely on gazetteers to link named places (e.g. Harlem) to map coordinates, the referential web of geography. NYPL has made an important contribution to the field with this project, building a gazetteer to create, verify, and connect data about New York City's places through time. From the early Lenape (a Native American tribe) names to the skyscraper now being built at One World Trade Center, the *New York City Chronology of Place* will help resolve the vexing issue that place names, boundaries, and even natural features change over time. It also extends NYPL's work converting historical maps into data via a historical gazetteer, which over time will evolve into a comprehensive and authoritative reference work.

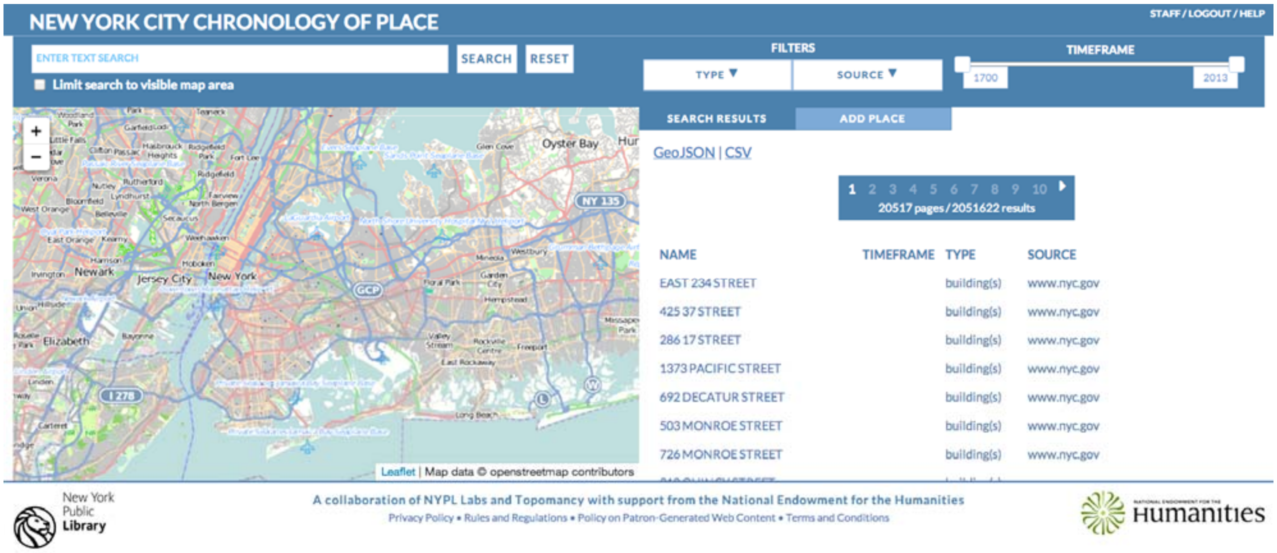
Project Genesis

Around 2006, the Library made an unsuccessful application to the NEH for a Research and Development grant aimed at creating a spatial indexing and display system for the collections digitized as part of NYPL's Digital Library Program. The underlying premise was that many collections were implicitly about places, but contained no explicit documentation of their location. In other words, metadata describing scanned photographs of the Grand Canyon was no more actionable as spatial than stills from the performance *The Grand Hotel*. In studying the comments made by the panel reviewers, it was clear that what NYPL needed to create a gazetteer to spatially organize its digital assets. Furthermore, this gazetteer needed to be populated with map information, more particularly historical maps, given the historical nature of the Library's collections and digitization tendencies.

Since the late 1990s, NYPL has digitized maps form its collection, but it wasn't until this realization in 2006 that the Library began to create historical map overlays using georectification, a cartographic technique long employed by practitioners of Geographic Information Science (GIS). This transformative process not only enabled researchers to study changing landscapes by geographically aligning new and old maps in digital space, but was also the first step in converting maps into machine readable data through tracing and annotation, enabling old maps information to be used as the basis for a gazetteer database.

Beginning in 2008, the NYPL, with contracted software developers, began developing a suite of tools at maps.nypl.org to transform the map collections into more useful derivatives, with the ultimate goal being to populate a yet-to-be-developed historical gazetteer. The first step in this transformation is map warping, whereby old maps are digitally stretched to geographically match the comparable location on a digital map, effectively normalizing images of maps across spatial coordinates. Once spatially aligned, the lines and boundaries defining geographic features on the map, like building footprints or rivers, can be traced and their attributes (names, feature type) transcribed. This unlocks potential future uses like keyword search (i.e. place name) across scanned maps, scaled linguistic analysis of those same names across numerous maps, and geocoding or integration into a gazetteer, ultimately enabling the visualization of historical, non-map digital objects like old photos or menus.

Having built these tools, the Library leveraged their potential, generating historical map data through the 2010 NEH-funded *New York City Historical GIS* project. Ultimately, some 9,000 maps were scanned, another 2,000 georectified, and close to 400 maps traced in their entirety. All of these are available through the NYPL's Digital Collections page and on maps.nypl.org. This project produced digital surrogates and derivatives which were useful in their own right, but the maps' information still needed a framework to display and integrate these results (a gazetteer) before it could be put to work. More specifically, the Library needed the means to connect and make explicit the relationships between geographic data from different timeframes. For example, the city of New Amsterdam became New York, a change reflected in the cartographic record, that can be made digitally evident through transformative work processes. Making these types of connections between troves of historical geographic data from multiple periods of time at scale effectively creates the fabric against which to perform many digital processes. These processes include the creation of new temporally conditioned authority files for historical places, to the organization and geotagging of troves of related, but not spatially specific metadata, and the geographic analysis of large body of text. This project, the *New York City Chronology of Place*, represents the Library's nascent foray into building a gazetteer.


NEW YORK CITY CHRONOLOGY OF PLACE

20517 pages / 2051622 results

NAME	TIMEFRAME	TYPE	SOURCE
EAST 234 STREET		building(s)	www.nyc.gov
425 37 STREET		building(s)	www.nyc.gov
286 17 STREET		building(s)	www.nyc.gov
1373 PACIFIC STREET		building(s)	www.nyc.gov
692 DECATUR STREET		building(s)	www.nyc.gov
503 MONROE STREET		building(s)	www.nyc.gov
726 MONROE STREET		building(s)	www.nyc.gov

A collaboration of NYPL Labs and Topomancy with support from the National Endowment for the Humanities

Privacy Policy • Rules and Regulations • Policy on Patron-Generated Web Content • Terms and Conditions

New York City Chronology of Place Search Interface

While the analytical processes remain aspirational, this project enabled NYPL to build the predicate technology necessary to make them possible, including the spatial/temporal database, basic search and indexing functions, more robust application programming interfaces (APIs), and authoring and editing tools.

Accomplishments

As noted in the interim report to NEH, NYPL learned at this project's outset (April 2012) that partner EntropyFreeLLP (subsequently reincorporated as Topomancy LLC) expected a delay in developing the web application and user interface. Topomancy was working on a grant with similar outcomes for the Library of Congress, and planned to repurpose parts of the database, the codebase for the applications programming interfaces, and the newly enhanced geocoder, as well as administrative and public user interfaces. The project start date was delayed forty-five days, extending its completion to mid-June 2013; consequently, NYPL requested and received a three-month-long project extension from the NEH.

At the project outset, the vendor built a fully versioned spatio-temporal database and index of gazetteer entries, representing geographic entities, including official name(s), identifier, feature type, locations, dates, and Linked Data references to equivalent representations (e.g. Geonames and OpenStreetmap among others). The initial data upload also included spatial data from maps.nypl.org such as 65,000 historical building footprints transcribed from *Maps of the City of New York* by William Perris, published in 1852, and a number of additional sources including the National Register of Historic Places, New York City's Landmarks Preservation Commission. Further data sources, such as the NYC Department of City Planning's Block and Lot boundaries and modern building footprints, were also included. This process was entirely collaborative, with NYPL providing appropriate prepackaged and cleaned sets of the required data to the developers and the developers building scripts for and conducting the "Extract, Transform and Load" (ETL) process necessary for populating a database. The database now contains more than 2 million spatial data objects, many to be de-duplicated through concordance as NYPL gets to work using the authoring tools described below.

In the next phase, the group developed an administrative interface for authoring and editing methods of the API for all classes of data in the *NYC Chronology of Place* and acceptance and rejection of submitted changes. The workplan also called for creation of an administrative interface for the maintenance of authoring and reviewing credentials. It soon became apparent that this task would be better suited for a later date when an editorial structure and critical mass of users is identified and invested in the project. The proposal also suggested that changes to gazetteer entries would enter into a quality assurance queue for accuracy checking, verification and confidence flagging, but what was actually built focused primarily on authoring and editing. This deviation from the workplan was necessitated by two factors. First, the work model involving multiple levels of editorial staff, where changes are queued, is a long term desire for this project, and a higher priority was assigned to solidifying the processes and interface components involved in the creation of gazetteer entries. Second, NYPL was not yet ready to commit the human resources necessary to build a fully fleshed gazetteer editorial staff.

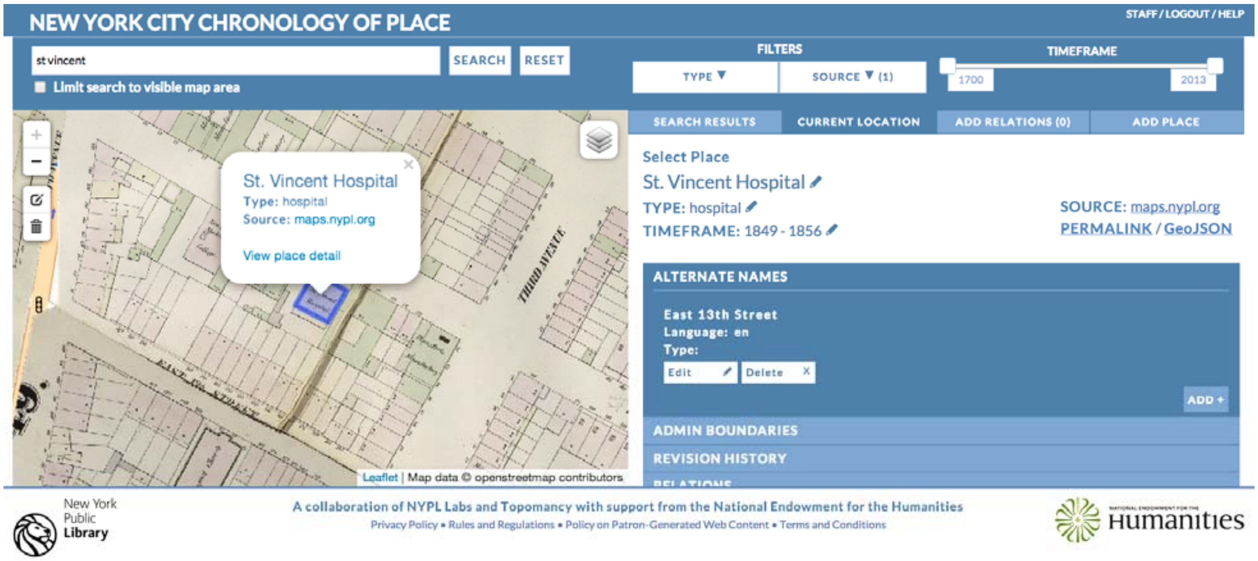
The administrative interface that was built enables users to edit existing gazetteer entries or add new entries features. This editorial functionality includes simple name changes and feature reclassification; extending date ranges (i.e. if a geographic feature such as Central Park was built in 1853, it's date range can be extended back to that time); addition of alternate names (i.e. if a feature is known by different names at different times, such as New York City and the Big Apple, this can be represented); or establishing relationships between features (i.e. Washington Parade Grounds "is replaced by" Washington Square Park). The interface to search, find, select and create relationships between multiple features proved complicated, requiring significant developer time and attention. Additionally, an interface to compare individual features and tie them to newer and older historical maps was created, allowing for the digital connection of features forward and backwards in time. For example, if a building footprint traced from an 1852 map is the same as one from 1922, a concordance can be made between the features, effectively merging them into one entry in the gazetteer, while maintaining references to the various historical map sources from which each were derived.

Throughout the grant period, in order to support the development of administrative and public user interfaces, the project team built a web services application programming interface (API) to enable searching for gazetteer entries by name, feature type, timeframe, and location. For simplicity and ease of use, the API was based on the REST model (http://en.wikipedia.org/wiki/Representational_state_transfer) and GeoJSON data format (http://geojson.org/), both well documented and understood, providing lightweight means for users of the data to develop client applications. For reasons noted above, the user interfaces for automatically pushing and recommending geographic feature changes to administrative users was not implemented, but the API services upon which this user interface can be built were created. Additionally, while the workplan called for the gazetteer API to be capable of conducting automated geographic remediation of metadata records, this capability has not yet been tested against the NYPL collections API because the Library's collections API was being built during the project period. This is one of the most important promises that this project has to offer, leading to the future ability of web developers to build geographic information retrieval applications for non-map collections.

One of the last user design components planned for this project was an authoring interface for users of maps.nypl.org to trace new gazetteer entries from georectified maps. After much discussion, it became clear that the gazetteer itself was the best place to author new entries for two reasons. First, the transcription interface in maps.nypl.org is itself in need of a user interface overhaul, currently an unfunded mandate. Additionally, some of the best and simplest-to-implement map-tracing tools share the software development framework (django) used to build *New York City Chronology of Place*. The lightweight, easy-to-use tracing tool that was adapted for this project was seamlessly integrated into the administrative user interface.

While the planned interface to allow NYPL to batch upload data from maps.nypl.org and elsewhere was not built, the ETL scripts and the API methods which enables NYPL technologists to upload these layers was well documented. The decision to not build this software is the same as noted above, namely that maps.nypl.org is several years old and in need of an upgrade. This decision was made as the Library was formulating a broader vision for working with maps on the web not as single user interface components but as pieces of a larger historical map platform. It did not make sense to connect the newer software to much older software at maps.nypl.org, understanding it might have a holistic makeover in the near future.

As the *NYC Chronology of Place* public interface now stands, gazetteer entries are accessible via geographic, temporal, and text search in the browsing interface, with initial entry through a zoomable map of New York City. Users can zoom to study areas or enter a keyword search, e.g. "hospital". The interface shows results in list format to the side and as a "zoomed to" location on the map. When multiple entries are located, they will show as "pinned" to the map or as "overlaid" transparent shapes. Clicking results will open a window with relevant information: data provenance, temporal frame, multilingual variants and alternate spellings. Multiple filters help cull results: a time slider to set or narrow search or results time frames, time proximity search helps find features before or after in time, and geographic proximity filtering searches for features within the selected pan and zoom map. Features within the user specified scope can be visualized on the map and in a results list to the side with a uri/url for each on the feature pages. Once searched and filtered, users can export data as geojson or csv file formats. Ultimately the API can publish web services (wfs, wfs-g/wgs, georss, rdf-xml and other formats) or as download (shp, kml, and other formats), but in the interest of simplicity, the decision was made to limit the export function in order to focus developer time on the user experience.


NEW YORK CITY CHRONOLOGY OF PLACE
STAFF / LOGOUT / HELP
SEARCH
RESET
FILTERS
TYPE ▼
SOURCE ▼ (1)
TIMEFRAME
SEARCH RESULTS
CURRENT LOCATION
ADD RELATIONS (0)
ADD PLACE
THIRD AVENUE
EAST 13th STREET
View place detail
Edit
Delete ×
ADD +
ADMIN BOUNDARIES
REVISION HISTORY
RELATIONS
Leaflet | Map data © openstreetmap contributors
New York Public Library
Humanities
NATIONAL ENDOWMENT FOR THE
A collaboration of NYPL Labs and Topomancy with support from the National Endowment for the Humanities
Privacy Policy • Rules and Regulations • Policy on Patron-Generated Web Content • Terms and Conditions

st vincent

□ Limit search to visible map area

TIMEFRAME: 1700 – 2013

St. Vincent Hospital
Type: hospital
Source: maps.nypl.org

Select Place

St. Vincent Hospital

TYPE: hospital

TIMEFRAME: 1849 - 1856

SOURCE: maps.nypl.org

PERMALINK / GeoJSON

ALTERNATE NAMES

East 13th Street
Language: en
Type:

Place Detail Page - St. Vincent Hospital

Another functionality, the "search within feature”, a query for gazetteer entries falling within the geographic boundary of a selected item, was scoped but not built. While important, it was less of a priority than the general search and editing user experience.

In summary, the broad project goals were successfully met: building a spatial and temporal database from diverse sources; implementation of a robust application programming interface (API) upon which to build user interfaces into the data and to enable the gazetteer to be used in other spatial organizing tasks; development of administrative and editor interfaces to add, transform, connect and delete gazetteer entries; and the roll-out of a public user interface for advanced querying of the gazetteer data. The Library, with our technology partners, utilized an iterative software development framework that made it possible for the project to unfold organically and flexibly. In some cases, this meant more time was spent honing fundamental aspects of the work, such as feature editing and search, and in other cases tasks beyond the immediate project scope, such as socially driven editorial features and the gazetteer/warper integration were postponed.

Audiences

On June 4 and 5, 2013, NYPL hosted a two-day hackathon, *Mobilizing Historic Geodata: Hack NYC's Past* with the primary goal of utilizing the Library’s historical spatial information and *NYC Chronology of Place* to organize and present other types of non-map materials by space and time. For the hackathon, the Library seeded several ideas for new computer applications, ideally public-facing, that would leverage the data from the *NYC Historical GIS* and the infrastructure (API) and database of NYC Chronology of Place.

One idea considered was *Historical Time Travel Receipt*, a mobile check-in application similar to Foursquare but for historical information. The premise is that with a GPS enabled smartphone, a user could arrive at a location, (i.e. NYPL or Bryant Park), and “check-in” to leave a digital footprint. The “check-in” location would prompt queries of that place (latitude/longitude), returning old maps from our collections. Additionally, it would query non-map databases using historical names associated with the location. A user who “checks-in” at Bryant Park would see historical maps of the area, gazetteer entries for places that once occupied the site of Bryant Park (like the Crystal Palace), as well as newspaper articles from *Chronicling America* and the historical New York Times, images from the NYPL Repository, entries from Wikipedia, and other alternate representations of that place found on the web.

One hacker team took a tangible step toward creating a more fully-realized *Historical Time Travel Receipt* by building a fully-functioning application called NYPL Time Traveller. The application uses the Foursquare programming interface to enable users to check in on mobile phones at any location in New York City and access the Library’s collection of more than 25,000 historical building photos.

Another very exciting development from the hackathon was the creation of the “Vectorizer” tool, created by NYPL Labs, which uses computer vision algorithms to automatically trace geographic features from old maps, speeding the process by which we gather actionable digital information from old maps. This data is now fed through the wildly popular crowdsourcing tool available at buildinginspector.nypl.org, which was created this summer after the close of the grant project.

While these projects do not use the *New York City Chronology of Place* APIs or the user interfaces, there were several very important outcomes from the hackathon. First, the 40 or so attendees included academics, librarians, public historians, hackers, journalists and others. This two-day event was an important point of inflection for the Library, and marks the first time a group organized around a specific set of collections and technologies has assembled. It also gestures to a clarification of the roles libraries can and should play for society as they evolve. Next, through the projects realized, NYPL helped to raise awareness within the New York City hacking community, a community which likely was previously unaware of the depth of NYPL’s historical, place-based collections, and had perhaps not thought deeply about the challenges ahead in uncovering such collections in a networked environment.

Finally, through the development of the “Vectorizer” tool, which led to the deployment of the wildly popular buildinginspector.nypl.org, the Library has managed to articulate one of the means by which we can scale the production of data that will be imported into the *New York City Chronology of Place*. By widening this bottleneck, NYPL is now able to focus on the issues that still remain at hand, namely, the reworking of the Map Warper and tracing tools and the integration of these into the gazetteer.

Through the course of this project, NYPL uncovered new needs as well as weaknesses in the toolchain, suggesting a reworking and updating of this component. One previously mentioned example is the somewhat outdated tracing tool on the Map Warper. Another need highlighted by the development process is the need for a better integration between the Map

Warper/tracing tool and the gazetteer. Taking this and other needs into consideration and with a broader community of practice in mind, NYPL has applied for a small grant from the Alfred P. Sloan Foundation to support hosting a meeting to establish a research and development agenda for historical maps on the web.

Evaluation

The number of unanticipated and unsolicited invitations received by the Project Director Matt Knutzen to speak and give interviews indicate the project was very well-received and impactful. The major success of the project is that the Library developed a new database, search interface and API for organizing collections based on space and time. Ultimately, this will result in the development of new free and open sourced software tools for libraries and the creative use and continued relevancy of NYPL's collections beyond this project.

The project's weaknesses include what might be characterized as an ambitious workplan, which included many advanced features one might expect from more mature software. The actual process of uncovering these weaknesses and the flexibility of the team in making adjustments demonstrates the strength of the software development philosophy the NYPL is now utilizing. The team made agile adjustments to achieve great strides in the application and learn about the sequencing and staging of database and interface components.

Continuation of the Project

The *New York City Chronology of Place* was originally conceived as experimental project, but based on its success, Library is committed to continuing these activities. In order to do so, NYPL plans to submit an application for an NEH Digital Humanities Implementation grant, to support the following tasks:

- Build a pipeline and framework for the mass scale load in, cleaning, organizing of historic geodata from a community of practice (e.g. librarians, hackers, civic data wranglers/producers, wikipedians, historians, journalists, geographers, local historians, genealogists, real estate developers, archaeologists) or anyone who might have an interest in collectively building and organizing resources around historical geography;

- Create an optimized search tool, first by conducting usability analysis and then rewriting the current gazetteer framework. Then build a named entity tagger for historical places (a means to identify, highlight and extract places from texts). This will help the Library build more robust, composite representations of historical places. For this we'd like to experiment with some pretty cutting edge natural language processing and machine learning approaches;

- And finally, leverage the gazetteer technology within the architecture of NYPL's data ecosystem, while at the same time providing that code, data and service to anybody who might find it useful, such as other libraries and archives or companies involved in spatial search like FourSquare and Google.

Long Term Impact

While the *New York City Chronology of Place* represents newly opened opportunities for serendipitous discovery, its true promise is its role as a spatial and temporal glue that connects map and spatial information to non-map digital information about those same places. The gazetteer is the predicate for doing this. It is the means through which the Library's entry for the Crystal Palace (present-day Bryant Park in New York City) can be programmatically linked to various newspaper articles from the joint NEH/LC *Chronicling America* historical newspaper project. The gazetteer will also eventually connect historical locations and maps to the NYPL's extensive digital collections portal, preservation repository, transcribed menus, playbills and theater programs, and animated stereographic photographs. It is the Library's long term goal to fulfill this promise by scaling the project's functionality to encourage vetted external groups to both contribute and edit data; continuing to build popular crowdsourcing tools to work data into usable formats for ingestion into the gazetteer; building the means by which data can be not only produced but woven together; modernizing the Map Warper technology toolchain so that it seamlessly integrates into the gazetteer; and leveraging the gazetteer technology against other NYPL assets and web properties.